Speaker Weight Estimation from Speech Signals Using a Fusion of the i-vector and NFA Frameworks

Amir Hossein Poorjam*, Mohamad Hasan Bahari* and Hugo Van hamme*
* Center for Processing Speech and Images, KU Leuven, Belgium
Ah.poorjam@iaukhsh.ac.ir , MohamadHasan.Bahari@esat.kuleuven.be , Hugo.Vanhamme@esat.kuleuven.be

***Abstract*— In this paper, a novel approach for automatic speaker weight estimation from spontaneous telephone speech signals is proposed. In this method, each utterance is modeled using the i-vector framework which is based on the factor analysis on Gaussian Mixture Model (GMM) mean supervectors, and the Non-negative Factor Analysis (NFA) framework which is based on a constrained factor analysis on GMM weights. Then, the available information in both Gaussian means and Gaussian weights is exploited through a feature-level fusion of the i-vectors and the NFA vectors. Finally, a least-squares support vector regression (LS-SVR) is employed to estimate the weight of speakers from given utterances.**

The proposed approach is evaluated on the telephone speech signals of National Institute of Standards and Technology (NIST) 2008 and 2010 Speaker Recognition Evaluation (SRE) corpora. Experimental results over 2339 utterances show that the correlation coefficients between actual and estimated weights of male and female speakers are 0.56 and 0.49, respectively, which indicate the effectiveness of the proposed method in speaker weight estimation.

***Index Terms*— i-vector, Non-negative Factor Analysis, Least-Squares Support Vector Regression, Speaker Weight Estimation.**

I. INTRODUCTION

The voice of a speaker conveys information about speaker's traits and states such as age, gender, body size (weight/height) and emotional state. Weight is a long term trait of a speaker which is considered as an important parameter in various applications. Speaker weight estimation is an interesting and challenging task in forensic, medical and commercial applications. In forensic scenarios, body size estimation of suspects from their voices can direct investigations to find cues in judicial cases. In service customization, automatic weight estimation may help users to receive services proportional to their physical conditions.

The relation between the size of various components of the sound production system (such as vocal folds and vocal tract) and the body size of a speaker has motivated researchers in the field of speaker recognition to look for features of an acoustic signal that provide cues to the body size of speakers. For instance, authors in [1] found a relationship between formants and the length of the vocal tract, based on the source-filter theory. Thus, since the vocal tract is a part of speaker's body, this feature can be used to estimate the weight of a speaker [2].

However, speaker weight estimation from the voice patterns is challenging. For instance, mean fundamental frequency (f_0) of voice is reported as a feature which has a (negative) correlation with body size. That is, females and children have higher f_0, while in males (who are taller and heavier), this value is lower [3]. However, when the relation of the fundamental frequency (f_0) and weight was investigated within male and female speakers, no correlation was found between f_0 and the weight of adult humans [4,5].

The lowest fundamental frequency of voice (F_0^{min}) is another feature which is determined by the mass and length of the vocal folds [3]. By investigating this feature, researchers have found no correlation between F_0^{min} and weight in adult human speakers [4,5].

Fitch has found formant dispersion (the averaged difference between adjacent pair of formant frequencies) a reliable feature which has a correlation with both vocal tract length and body size in macaques [6]. However, a weak relation between formant parameters and weight of human adults is reported in study conducted by Gonzalez [7]. This weak correlation may be due to the fact that the vocal folds in humans at puberty grow independent of the rest of the head and body. This issue is more evident in the males than the females [8, 9].

Gonzalez studied the correlation between formant frequencies and weight in human adults [7]. He calculated the formant parameters by means of a long-term average analysis of running speech signals uttered by 91 speakers. In this experiment, the Pearson correlation coefficients between formants and weights for male and female speakers were reported to be 0.33 and 0.34, respectively [7].

In research conducted by Van Dommelen and Moxness [10], the ability to judge the weight of speakers from their speech samples was investigated. They reported a significant correlation between estimated weight (judged by listeners) and actual weight of only male speakers. In addition, they performed a regression analysis involving several acoustic features such as f_0, formant frequencies, energy below 1 kHz, and speech rate. The results showed that the speech rate was the only parameter which had a significant correlation with male speaker's weight. They concluded that speech rate of male speakers is a reliable predictor for weight estimation.

Modeling speech utterances with Gaussian mixture model (GMM) mean supervectors is demonstrated to be an effective approach to speaker recognition [11]. However, GMM mean supervectors are high dimensional vectors, and obtaining a reliable model is difficult when limited data are available.

Recently, utterance modeling using the i-vector framework [12] has considerably increased the accuracy of the classification and regression problems in the field of speaker characterization [13–15]. The i-vector, which is based on the factor analysis on GMM mean supervectors, represents an utterance in a compact and a low-dimensional feature vector. In addition, various studies show that although GMM weights convey less information than GMM means, they provide complementary information to GMM means [16–18]. A Non-negative Factor Analysis (NFA) framework [16] which is based on a constrained factor analysis for GMM weights, has been recently introduced and yields a new low-dimensional utterance representation.

In this study, a new speech-based method for automatic weight estimation is proposed. In this approach, instead of using raw acoustic features, each utterance is modeled using the i-vector and the NFA frameworks. Then, through a feature-level fusion of the i-vectors and the NFA vectors, the available information in both Gaussian means and Gaussian weights is exploited to enhance the accuracy of automatic speaker weight estimation. To perform function approximation, a least-squares support vector regression (LS-SVR) is utilized, and the effect of the kernel in LS-SVR is investigated. The proposed method is evaluated on spontaneous telephone speech signals of the NIST 2008 and 2010 SRE corpora. Experimental results confirm the effectiveness of the proposed approach.

The rest of the paper is organized as follows. The problem of automatic weight estimation is formulated and the proposed approach is described in Section II. Section III explains the experimental setup. The evaluation results are presented and discussed in Section IV. The paper ends with conclusions in Section V.

II. System Description

In this section, the problem of automatic weight estimation is formulated and the main constituents of the proposed method are described.

A. Problem Formulation

In the speaker weight estimation problem, we are given a set of training data $D = \{O_i, y_i\}_{i=1}^{N}$, where O_i denotes the i^{th} utterance and $y_i \in \mathbb{R}$ denotes the corresponding weight.

The goal is to approximate a function g, such that for an utterance of an unseen speaker, O_{tst}, the estimated weight, $\hat{y}=g(O_{tst})$, approximates the actual weight as good as possible.

B. Utterance Modeling

By fitting a GMM to acoustic features extracted from each speech signal, a variable-duration speech signal is converted into a fixed-dimensional vector which is suitable for regression algorithms. The parameters of the obtained GMM characterize the corresponding utterance. Due to limited data, we are not able to accurately fit a separate GMM for a short utterance, specially in the case of GMMs with a high number of Gaussian components. Thus, for adapting a universal background model (UBM) to characteristics of utterances in training and testing databases, parametric utterance adaptation techniques are applied. In this paper, the i-vector and the NFA frameworks are applied to adapt UBM means and weights, respectively.

1) Universal Background Model and Adaptation:

Consider a UBM with the following likelihood function of data $O=\{\mathbf{o}_1, \ldots, \mathbf{o}_t, \ldots, \mathbf{o}_T\}$.

$$p(\mathbf{o}_t \mid \gamma) = \sum_{c=1}^{C} \pi_c \, p(\mathbf{o}_t \mid \mu_c, \mathbf{\Sigma}_c)$$
$$\gamma = \{\pi_c, \mu_c, \mathbf{\Sigma}_c\} \quad , \quad c = 1, \ldots, C \qquad (1)$$

where $\mathbf{o}_t$ is the acoustic vector at time t, π_c is the mixture weight for the c^{th} mixture component, $p(\mathbf{o}_t|\mu_c, \mathbf{\Sigma}_c)$ is a Gaussian probability density function with mean μ_c and covariance matrix $\mathbf{\Sigma}_c$, and C is the total number of Gaussian components in the mixture. The parameters of the UBM –γ– are estimated on a large amount of training data.

2) The i-vector Framework:

One effective method for speaker weight estimation involves adapting UBM means to the speech characteristics of the utterance. Then, the adapted GMM means are extracted and concatenated to form Gaussian mean supervectors. However, since Gaussian components of the UBM model are adapted independent of each other, some components are not updated in the case of limited training samples [19]. This problem can be alleviated by linking the Gaussian components together using the Joint Factor Analysis (JFA) framework [20].

In the JFA framework, each utterance is represented by a supervector $\mathbf{M}$ which is a speaker- and channel-dependent vector of dimension $(C.F)$, where C is the total number of the mixture components in a feature space of dimension F. In the JFA framework, it is assumed that $\mathbf{M}$ can be decomposed into two supervectors:

$$\mathbf{M} = \mathbf{s} + \mathbf{c} \qquad (2)$$

where $\mathbf{s} = \mathbf{u} + \mathbf{V}\boldsymbol{q} + \mathbf{D}\boldsymbol{r}$ is a speaker-dependent supervector and $\mathbf{c} = \mathbf{U}\boldsymbol{p}$ is a channel-dependent supervector. $\mathbf{s}$ and $\mathbf{c}$ are independent and possess normal distributions. $\mathbf{u}$ is the speaker- and channel-independent supervector, $\mathbf{V}$ defines a lower dimensional speaker subspace, $\mathbf{U}$ is a lower dimensional channel subspace, and $\mathbf{D}$ defines a speaker subspace. $\boldsymbol{q}$ and $\boldsymbol{r}$ are factors in speaker subspace, and $\boldsymbol{p}$ is a channel-dependent factor in channel subspace. The vectors $\boldsymbol{p}$, $\boldsymbol{q}$ and $\boldsymbol{r}$ are random variables with standard normal distributions $N(0, I)$ which are jointly estimated.

In the JFA framework, the channel factor contains some information about speakers, which can be utilized in speaker identification. This fact resulted in proposing a new utterance modeling approach, referred to as the i-vector framework or the total variability modeling [21]. This method comprises both speaker variability and channel variability. Channel compensation procedures such as within-class covariance normalization (WCCN) can be further applied to compensate the residual channel effects in the speaker factor space [22].

The i-vector framework assumes that each utterance possesses a speaker- and channel-dependent GMM supervector which its mean, $\mathbf{M}$, can be decomposed as

$$\mathbf{M} = \mathbf{u} + \mathbf{T}\mathbf{v} \qquad (3)$$

where **u** is the mean supervector of the UBM, and **T** spans a low-dimensional subspace (400 dimensions in this work). In the i-vector framework, **T** and **v** are estimated using the Expectation-Maximization (EM) algorithm. In the E-step, **T** is supposed to be known, and **v** is updated. In the M-step, **v** is assumed to be known, and **T** is updated. The subspace vector **v** is treated as a hidden variable with the standard normal prior and the i-vector is its maximum-a-posteriori (MAP) point estimate which is calculated by maximization of the following auxiliary function over **v**.

$$\Psi(\gamma,\mathbf{v})=\sum_{t=1}^{T}\sum_{c=1}^{C}\theta_{c,t}\log\pi_c p(\mathbf{o}_t\mid[\mu_c+\mathbf{T}_c\mathbf{v}],\mathbf{\Sigma}_c)N(\mathbf{v}) \quad (4)$$

where $N(\mathbf{v})$ denotes the standard normal distribution of **v**, $\mathbf{T}_c$ are the rows of the subspace matrix **T**, which correspond to the c^{th} Gaussian mean, and $\theta_{c,t}$ is the occupation count for the c^{th} mixture component and t^{th} frame. The occupation count is calculated as follows:

$$\theta_{c,t}=\frac{\pi_c p(\mathbf{o}_t\mid\mu_c,\mathbf{\Sigma}_c)}{\sum_{c=1}^{C}\pi_c p(\mathbf{o}_t\mid\mu_c,\mathbf{\Sigma}_c)} \quad (5)$$

In the E-step, the posterior distribution of **v** is Gaussian with the following mean $\mathbf{v}_\mu$ and covariance matrices $\mathbf{v}_\sigma$ [23]:

$$\mathbf{v}_\sigma=\left[\mathbf{I}+\sum_c\theta_c\mathbf{T}_c'\overline{\mathbf{\Sigma}}_c^{-1}\mathbf{T}_c\right]^{-1} \quad (6)$$

$$\mathbf{v}_\mu=\mathbf{v}_\sigma\sum_c\left[\mathbf{T}_c'\overline{\mathbf{\Sigma}}_c^{-1}\sum_t\theta_{c,t}(\mathbf{o}_t-\mathbf{m}_c)\right] \quad (7)$$

where **I** denotes an identity matrix of appropriate size, $\mathbf{m}_c$ and $\overline{\mathbf{\Sigma}}_c$ are adapted mean and covariance of the c^{th} Gaussian, which are updated during each EM iteration starting from UBM parameters, and $'$ represents the transpose operator.

In the M-step, the subspace matrix **T** is estimated via maximization of the following auxiliary function over **T**.

$$\tilde{\Psi}(\gamma,\mathbf{T})=\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{c=1}^{C}\theta_{c,t,i}\log\pi_{c,i}p(\mathbf{o}_{t,i}\mid[\mu_c+\mathbf{T}_c\mathbf{v}_i],\mathbf{\Sigma}_{c,i}) \quad (8)$$

An efficient procedure for training **T** and for MAP adaptation of the i-vectors can be found in [23].

In the total variability modeling approach, the i-vector is the low-dimensional representation of an audio recording that can be used for classification and estimation purposes.

3) The Non-negative Factor Analysis (NFA) Framework:

The NFA is a new framework for adaptation and decomposition of GMM weights based on a constrained factor analysis [16]. The basic assumption of this method is that for a given utterance, the adapted GMM weight supervector can be decomposed as follows:

$$\mathbf{w}=\boldsymbol{\pi}+\mathbf{L}\mathbf{r}, \quad (9)$$

where $\boldsymbol{\pi}$ is the UBM weight supervector (2048 dimensional vector in this study). **L** is a matrix of dimension $C\times\rho$ spanning a low-dimensional subspace (300 dimensions in this work). **r** is a low-dimensional subspace vector obtained through a constrained maximum likelihood estimation criterion.

In this framework, the adapted weights are obtained by maximizing the following objective function over w_c .

$$\Psi(\gamma,\mathbf{r})=\sum_{t=1}^{T}\sum_{c=1}^{C}\theta_{c,t}\log w_c p(\mathbf{o}_t\mid\mu_c,\mathbf{\Sigma}_c) \quad (10)$$

Substituting w_c by $(\pi_c+\mathbf{L}_c\mathbf{r})$ in the Eq. 10, and given an utterance *O*, a maximum likelihood estimation of **r** is obtained by solving the following constrained optimization problem:

$$\max_{\mathbf{r}}\left(\Psi(\gamma,\mathbf{r})\right)=\max_{\mathbf{r}}\left(\bar{\theta}'(O)\log(\boldsymbol{\pi}+\mathbf{L}\mathbf{r})\right) \quad (11)$$

$$\text{Subject to}\begin{cases}\mathbf{1}(\boldsymbol{\pi}+\mathbf{L}\mathbf{r})=1\\ \boldsymbol{\pi}+\mathbf{L}\mathbf{r}>0\end{cases}$$

where **1** is a row vector of dimension *C* with all elements equal to one, and $\bar{\theta}(O)=\sum_t[\theta_{1,t}\ \ \ldots\ \ \theta_{C,t}]'$.

In this framework, neither subspace matrix **L** nor subspace vector **r** are constrained to be non-negative. However, unlike the i-vector framework, the applied factor analysis for estimating the subspace matrix **L** and the subspace vector **r** is constrained such that the adapted GMM weights are non-negative and sum up to one. The procedure of calculating **L** and **r** involves a two-stage algorithm similar to EM and can be found in [16]. The subspace matrix **L** is estimated over a large training dataset. It is then used to extract a subspace vector **r** for each utterance in train and test datasets.

This new low-dimensional utterance representation approach was successfully applied to speaker characterization [17, 24] and language/dialect recognition [16] tasks.

4) Feature-Level Fusion of the i-vector and the NFA Frameworks:

Previous studies show that although GMM weight supervectors contain less information than GMM means, they provide complementary information to GMM means [18]. Feature-level fusion and score-level fusion are considered as effective approaches to exploit available information in both GMM means and weights [18, 24]. Score-level fusion in which the outputs of different estimators are fused, requires a development data set to train the fusion model, which results in decreasing the number of training data. However, fusion at feature level in which various features are normalized and concatenated, eliminates the need for assigning a considerable amount of available training data for development set, and estimation can be performed in one learning phase.

In this paper, a feature-level fusion of the i-vectors and the NFA vectors is considered to improve the estimation accuracy. As illustrated in Fig. 1, the i-vectors and the NFA vectors should be normalized prior to concatenation. To this aim, extracted i-vectors and the NFA vectors are mapped into a low-dimensional space using linear discriminant analysis (LDA). Then, the obtained low-dimensional vectors are concatenated to form a longer vector.

C. Function Approximation

In this study, a least squares support vector regression (LS-SVR) is employed to estimate speaker weight.

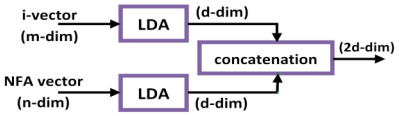

Fig. 1. Block diagram of the utterance modeling in feature-level fusion.

1) Least Squares Support Vector Regression:

Support vector regression (SVR) is a function approximation approach developed as a regression version of the widely known Support Vector Machines (SVM) classifier. Using nonlinear transformations, SVMs map the input data into a higher dimensional space in which a linear solution can be calculated. They also keep a subset of the samples which are the most relevant data for the solution and discard the rest. This makes the solution as sparse as possible. While SVMs perform the classification task by determining the maximum margin separation hyperplane between classes, SVR carries out the regression task by finding the optimal regression hyperplane in which most of training samples lie within an ε-margin around this hyperplane [25].

In this study, we use the least squares version of support vector regression (LS-SVR). While an SVR solves a quadratic programming with linear inequality constraints, which results in high algorithmic complexity and memory requirement, an LS-SVR involves solving a set of linear equations by considering equality constraints instead of inequalities for classical SVR [25], which speeds up the calculations. This simplicity is achieved at the expense of loss of sparseness. Therefore, all samples contribute to the model, and consequently, the model often becomes unnecessarily large.

In order to investigate the effect of the kernel in LS-SVR, linear and radial basis function (RBF) kernels are used. For the LS-SVR with RBF kernels, a K-fold cross-validation is used to tune the smoothing parameter of the kernels.

D. Training and Testing

The proposed weight estimation approach is illustrated in Fig. 2. During the training phase, each utterance in the training data set is mapped to a high dimensional vector using one of the mentioned utterance modeling approaches described in Section II-B. Then, the obtained vectors along with their corresponding weight labels are used to train an estimator for approximating function g.

During the testing phase, the same utterance modeling approach applied in training phase is used to extract a high dimensional vector from a test utterance. Then, the estimated weight is obtained using the trained regression function.

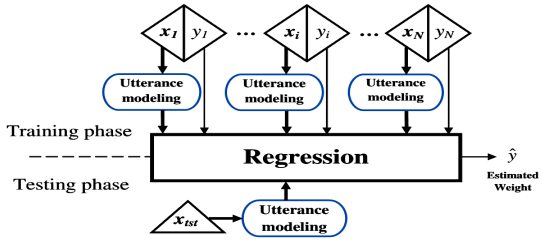

Fig. 2. Block diagram of the proposed speaker weight estimation approach in training and testing phases.

III. Experimental Setup

A. Database

The National Institute for Standard and Technology (NIST) have held annual or biannual speaker recognition evaluations (SRE) for the past two decades. With each SRE, a large corpus of telephone conversations are released. Conversations typically last 5 minutes and originate from a large number of speakers for whom additional meta data (such as age, height, weight, language and smoking habits) is recorded.

The NIST databases were chosen for this work due to the large number of speakers and because the total variability subspace requires a considerable amount of development data for training. The development data set used to train the total variability subspace and UBM includes over 30,000 speech recordings and was sourced from the NIST 2004-2006 SRE databases, LDC releases of Switchboard 2 phase III and Switchboard Cellular (parts 1 and 2).

For the purpose of automatic speaker weight estimation, telephone recordings from the common protocols of the recent NIST 2008 and 2010 SRE databases are pooled together to create a dataset of 8241 utterances uttered by 1333 speakers. Then, it is divided into two disjoint parts such that 80% and 20% of all speakers are used for training and testing sets, respectively. Thus, of all 8241 utterances, 5902 utterances are considered for training set and 2339 utterances are considered for testing set. Fig.3 shows the weight histograms of training and testing datasets for male and female speakers.

B. Performance Metric

In order to evaluate the effectiveness of the proposed method, the mean-absolute-error (MAE) of the estimated weight, and the Pearson correlation coefficient (*CC*) between the actual and estimated weights are used. MAE is defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i - y_i\right| \qquad (12)$$

where $\hat{y}_i$ is the i^{th} estimated weight, y_i is the i^{th} actual weight, and N is the total number of test samples.

The Pearson correlation coefficient is computed as:

$$CC = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{\hat{y}_i - \mu_{\hat{y}}}{\sigma_{\hat{y}}}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right) \qquad (13)$$

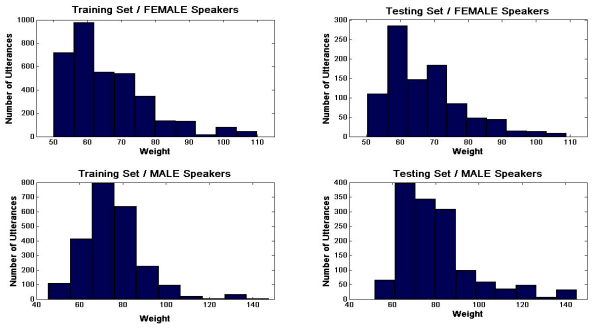

Fig. 3. The weight histograms of telephone speech utterances of training and testing datasets for male and female speakers.

where μ_y and σ_y denote the mean and standard deviation of the actual speakers' weight respectively, and $\mu_{\hat{y}}$ and $\sigma_{\hat{y}}$ are the mean and standard deviation of the estimated weights respectively.

IV. Results and Discussion

In this section, the proposed speaker weight estimation approach is evaluated. The acoustic feature vector is a 60-dimensional vector consists of 20 Mel-Frequency Cepstrum Coefficients (MFCCs) including energy appended with their first and second order derivatives. MFCCs are obtained using cosine transform of the real logarithm of the short-term energy spectrum represented on a mel-frequency scale [26]. This type of feature is very common in the i-vector-based speaker recognition systems. Wiener filtering, feature warping [26] and voice activity detection [27] have also been considered in the front-end processing to obtain more reliable features.

In this study, an LS-SVR has been employed to perform weight estimation. To evaluate the effect of the kernel in LS-SVR, two different kernels, namely linear and radial basis function (RBF) kernels are used. The hyper-parameters of the RBF kernel are tuned using a 15-fold cross-validation. After optimization of the hyper parameters, the model is trained. The LS-SVR models are implemented using LS-SVMlab1.8 Toolbox [28] in Matlab environment.

To investigate the effect of applied feature-level fusion on automatic weight estimation and to evaluate the effectiveness the proposed speaker weight estimation approach, it is worth comparing the proposed method with two systems, namely the basic estimator and the i-vector-based system.

A. The Basic Estimation System

When an utterance of an unseen speaker is applied to a basic estimator, its output is the average weight of speakers in training data set. The basic estimation system provides us a chance level accuracy. The results of using a basic estimator for speaker weight estimation are reported in the first row of Table I. Besides providing a reference level for speaker weight estimation systems, the basic estimator highlights a limitation of using mean-absolute-error as a performance metric for weight estimation problem. The MAE is limited in some respects, specially, in the case of a test set with a skewed distribution which is the case in this task. When a test data set with a skewed distribution is applied to a basic estimator, the MAE might be in an acceptable range, based on the variance of the data. For instance, when the database described in Section III-A was applied to the basic estimator, the MAE for male and female speakers were 12.93 kg and 9.03 kg, respectively. However, the measured *CC* for males and females were equal to zero. For this reason, the correlation coefficient is a preferred performance metric in this task, which reflects the performance of the estimators in a more sensible way.

B. The i-vector-based System

In the i-vector-based system, each utterance in training set is mapped to a low-dimensional vector (400 dimensions in this work) using the i-vector framework. Then, the extracted i-vectors along with their corresponding weight labels are used to train estimator. The results of employing an LS-SVR as an estimator, and using the i-vector framework for utterance modeling are presented in the second and the third rows of Table I. Comparison of the i-vector-based system with the basic estimator shows the effectiveness of the i-vectors in automatic speaker weight estimation. The obtained results also indicate that the linear kernel leads to a more accurate estimation compared with the RBF kernel. Thus, the LS-SVR with the linear kernel is selected for the rest of experiments.

C. The Proposed Weight Estimation Approach

To improve the estimation accuracy of the i-vector-based weight estimation, a feature-level fusion of the i-vectors and the NFA vectors is considered in this paper. In the proposed method, the extracted i-vectors and NFA vectors are normalized and concatenated to form a longer vector. The obtained vector, along with the corresponding weight label is then used to train estimator. The last row of Table I contains the results of the proposed weight estimation approach. The obtained results indicate that the accuracy of weight estimation increases after feature-level fusion compared with the estimation using the i-vector-based estimator. It concurs with the previous studies demonstrating that GMM weights provide complementary information to GMM means. The achieved relative improvements in *CC* by the proposed fusion scheme compared with the i-vector-based estimator for male and female speakers are 14.28% and 2.04%, respectively, which show that the proposed method is more effective in weight estimation for male speakers than for female speakers.

In [24] we proposed a multitask speaker characterization approach to simultaneously estimate age, weight and height of speakers from speech signals, based on a score-level fusion of the i-vector and the NFA frameworks. Comparing the results of these two fusion schemes reveal that fusion of the i-vector and the NFA frameworks at feature level is more effective in speaker weight estimation. In addition, fusion at feature level eliminates the need for assigning a considerable amount of training data for development set, and performs speaker weight estimation in one learning phase.

The reported *CC* for speaker weight estimation based on the formant parameters of the running speech signals uttered by 91 speakers are 0.33 and 0.34 for male and female speakers, respectively [7]. The results obtained from our proposed speaker weight estimation seem reasonable, considering the fact that the applied testing dataset in this study consists of spontaneous speech signals and the number of speakers in test set is considerably larger than that of in [7]. It can be concluded that automatic speaker weight estimation using a fusion of the i-vector and NFA frameworks is more efficient compared with estimation based on the raw acoustic features.

TABLE I. The MAE (in kg) and *CC* of the proposed speaker weight estimation, compared with the basic and i-vector-based systems.

Estimator	Feature	MALE		FEMALE	
		CC	**MAE**	***CC***	**MAE**
Basic Estimator	---	0	12.93	0	9.03
LS-SVR (RBF)	i-vector	0.46	11.42	0.47	7.80
LS-SVR (Linear)	i-vector	0.48	11.53	0.48	7.86
LS-SVR (Linear)	i-vector & NFA	**0.56**	**11.16**	**0.49**	**7.79**

V. Conclusion

In this paper a novel approach for automatic speaker weight estimation from spontaneous telephone speech signals was proposed. In this method, each utterance was modeled using a fusion of the i-vector and the NFA frameworks at feature level. Through this new utterance modeling approach, the available information in both GMM means and GMM weights was utilized. Then, an LS-SVR was employed to estimate the weight of a speaker from a given utterance. The proposed method was trained and tested on the telephone conversations of NIST 2008 and 2010 SRE corpora.

Evaluation results over 2339 utterances show that the correlation coefficients between actual and estimated weights of male and female speakers after feature-level fusion are 0.56 and 0.49, respectively, which indicate the effectiveness of the proposed method in automatic speaker weight estimation compared with estimation based on the raw acoustic features.

Utilizing information in Gaussian weights in conjunction with that of in Gaussian means through a fusion of the i-vectors and the NFA vectors resulted in achieving 14.28% and 2.04% relative improvements in *CC* compared with the i-vector-based weight estimation system.

References

[1] G. Fant, Acoustic Theory of Speech Production. The Hague: Mouton, 1960.

[2] N. J. Lass and M. Davis, "An investigation on speaker height and weight identification," Journal of the Acoustical Society of America, vol. 60, pp. 700–703, 1976.

[3] C. Darwin, The Descent of Man and Selection in Relation to Sex. London: Murray, 1871.

[4] N. J. Lass and W. S. Brown, "Correlational study of speakers heights, weights, body surface areas, and speaking fundamental frequencies.", Journal of the Acoustical Society of America, vol. 63, pp. 1218–1220, 1978.

[5] H. J. Kunzel, "How well does average fundamental frequency correlates with speaker height and weight?", Journal of Phonetica, vol. 46, pp. 117–125, 1989.

[6] T. W. Fitch, "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," Acoustical Society of America, vol. 102, pp. 1213–1222, 1997.

[7] J. Gonzalez, "Formant frequencies and body size of speaker: a weak relationship in adult humans," Journal of Phonetics, vol. 32, pp. 277–287, 2004.

[8] U. G. Goldstein, "An articulatory model for the vocal tracts of growing children." Ph.D. dissertation, Massachusetts Institute of Technology, 1980.

[9] V. E. Negus, The Comparative Anatomy and Physiology of the Larynx. Hafner, New York, 1949.

[10] W. A. Van Dommelen and B. H. Moxness, "Acoustic parameters in speaker height and weight identification: sex-specific behavior," Language and Speech, vol. 38, pp. 267–287, 1995.

[11] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," IEEE Signal Process. Letters, vol. 13, no. 5, pp. 308–311, 2006.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Frontend factor analysis for speaker verification," IEEE Trans. Audio, Speech, and Lang. Process., vol. 19, no. 4, pp. 788–798, 2011.

[13] M. H. Bahari, M. McLaren, H. Van hamme, and D. Van Leeuwen, "Age estimation from telephone speech using i-vectors," in Proc. Interspeech, 2012, pp. 506–509.

[14] M. H. Bahari, "Automatic speaker characterization: Automatic identification of gender, age, language and accent from speech signals," Ph.D. dissertation, KU Leuven – Faculty of Engineering Science, Belgium, May 2014.

[15] A. H. Poorjam, M. H. Bahari, V. Vasilakakis, and H. Van hamme, "Height estimation from speech signals using i-vectors and least-squares support vector regression," in Proc. 37th International Conference on Telecommunications and Signal Processing, Germany, 2014.

[16] M. H. Bahari, N. Dehak, H. Van hamme, L. Burget, A. Ali, and J. Glass, "Non-negative factor analysis of Gaussian mixture model weight adaptation for language and dialect recognition," Transactions on Audio, Speech, and Language Processing, vol. 22, no. 7, pp. 1117–1129, July 2014.

[17] A. H. Poorjam, "Speaker profiling for forensic applications," Master's thesis, KU Leuven – Faculty of Engineering Science, 2014.

[18] M. H. Bahari, R. Saeidi, H. Van hamme, and D. van Leeuwen, "Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech," in Proc. ICASSP2013, 2013, pp.7344-7348.

[19] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," IEEE Transaction on Speech and Audio Processing, vol. 13, no. 3, pp. 345–354, 2005.

[20] S. Shum, N. Dehak, R. Dehak, and J. Glass, "Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification," in Proc. Odyssey, 2010.

[21] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling: Application to speaker verification," Ph.D. dissertation, Ecole de Technologie Superieure de Montreal, Montreal, QC, Canada, 2009.

[22] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in Proc. Interspeech, vol. 4, no. 2.2, 2006.

[23] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," IEEE Trans. Audio, Speech, and Lang. Process., vol. 16, no. 5, pp. 980–988, 2008.

[24] A.H. Poorjam, M.H. Bahari, and H. Van hamme, "Multitask speaker profiling for estimating age, height, weight and smoking habits from spontaneous telephone speech signals", in proc. 4th International Conference on Computer and Knowledge Engineering, Iran, 2014.

[25] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, Least squares support vector machines. World Scientific, 2002.

[26] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," pp. 213–218, 2001.

[27] M. McLaren and D. van Leeuwen, "A simple and effective speech activity detection algorithm for telephone and microphone speech," in Proc. NIST SRE Workshop, 2011.

[28] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J. A. K. Suykens, "Ls-svmlab1.8 toolbox," http://www.esat.kuleuven.be/sista/lssvmlab/.